BPGA: an interactive Shiny application for basic population genetic analysis of genotype data

Joan Fibla
Facultat de Medicina, Universitat de Lleida. Av. Rovira Roure 80, 25199 Lleida.
Secció Biomedicina, Institut d'Estudis Ilerdencs. Plaça de la Catedral, s/n, 25002 Lleida.
Corresponding author: joan.fibla@udl.cat; joan.fibla@gmail.com
ORCID: https://orcid.org/0000-0002-3255-2227

Abstract

Background: Population structure and ancestry inference are routine in human genetics, yet remain inconvenient for non-experts because canonical tools (PLINK, GCTA, ADMIXTURE) require command-line expertise and careful data management.

Results: BPGA (Basic Population Genetic Analysis) is an open-source R/Shiny application that provides an interactive workflow for educational and exploratory population genetic analyses on datasets in PLINK binary format. BPGA executes LD-pruned PCA, genome-wide *Fst* scans, and ADMIXTURE clustering from a guided interface, and it can optionally merge user data with curated worldwide reference panels (1000 Genomes Project and Human Genome Diversity Project). The app produces publication-ready figures (PNG/HTML), interactive views (plotly/leaflet), and preserves logs for reproducibility.

Availability: Source code is available at https://github.com/jfibla/BPGA-a-Shiny-app-to-perform-Basic-Population-Genetic-Analysis under the Apache-2.0 license. A hosted demo is also available at http://15.188.54.171:3838/bpga_app/.

Conclusions: BPGA lowers the barrier between raw genotype data and interpretable population-genetic summaries for research and teaching, while keeping analyses transparent and reproducible.

Footnote: "Submitted to PCI Genomics for evaluation."

Introduction

Understanding population structure is foundational in human genetics and genomics. It underpins ancestry inference, quality control, and the mitigation of population stratification in association studies, while also providing a gateway to evolutionary thinking for students entering the field. The canonical toolchain—PLINK for data management and PCA, GCTA for genome-wide *Fst* and ADMIXTURE for maximum-likelihood estimation of ancestry proportions—has proven robust and widely adopted [1–4]. Yet this ecosystem remains challenging for newcomers: command-line literacy, environment configuration, careful file hygiene, and reproducible reporting are all non-trivial hurdles that can distract from core population-genetic concepts.

BPGA (Basic Population Genetic Analysis) was designed to reduce that friction. Built in R/Shiny [5], it wraps standard analyses behind a browser-based interface that accepts binary PLINK datasets, offers sensible defaults for autosome filtering, MAF thresholds, and LD pruning, and runs PCA, *Fst* and ADMIXTURE with a single click. Interactive visualizations (scatterplots, Manhattan-style tracks, and leaflet maps) and downloadable, publication-ready figures help learners connect algorithms to interpretable outputs. Optional merging with curated worldwide references from the 1000 Genomes Project and the Human Genome Diversity Project [6,7] enables students to situate their data—class projects, cohort subsets, or simulated genotypes—within a familiar global context.

From an educational standpoint, BPGA provides an on-ramp for graduate teaching: students can reproduce literature-grade analyses without first mastering a Unix workflow, reinforcing concepts such as LD pruning, interpretation of principal components, K selection via cross-validation, and the interpretation and limits of *Fst*. From a research standpoint, BPGA supports rapid prospection on real cohorts: quick ancestry checks, detection of outliers, exploratory *Fst* scans between user-defined groups, and preliminary ADMIXTURE runs that can later be formalized in scripted pipelines. The same interface thus serves both early-stage pedagogy and hypothesis-generating analyses in active projects.

In the remainder of the paper we (i) describe BPGA's data model and analysis pipeline; (ii) outline reference-panel integration and exportable visualizations; and (iii) present research, teaching-oriented use cases that align with common graduate-level learning outcomes in population genetics.

Methods

Implementation

BPGA is implemented in R and relies on the following libraries: shiny/shinyjs for the interface, data.table, dplyr, tidyr for data handling, ggplot2/ggh4x for static plots, plotly for interactivity, leaflet and leaflet.minicharts for maps, and htmlwidgets for export. Computational work is delegated to external binaries on the host system: PLINK (1.9), GCTA, and ADMIXTURE.

A per-session workspace stores intermediate PLINK files, logs, and figures under pca/, Fst/, admx/, and plots/, which can be downloaded as a single archive.

Software checklist

- Code: https://github.com/jfibla/BPGA-a-Shiny-app-to-perform-Basic-Population-Genetic-Analysis (Apache-2.0).
- Platform: R ≥ 4.3; packages listed in the repository DESCRIPTION/README.
- External tools: PLINK (1.9), GCTA, ADMIXTURE available on the system PATH or configured paths.
- Data: Public references derived from 1000 Genomes and HGDP; example datasets included.
- Outputs: PCA.png, ADM.png, box_ADM.png, *Fst* _interactive.html, map_ADM.png plus PLINK/GCTA logs.
- Releases/archival: https://doi.org/10.5281/zenodo.16953021.

Input and data model

BPGA accepts user datasets as compressed archives (.zip or .gz) containing PLINK binary files (.bed/.bim/.fam).

FAM population coding (first column, FID) — choose one when loading:

- POP1_POP2 (“pop1pop2”): if samples belong to several superpopulations and subpopulations, concatenate the superpopulation and subpopulation codes (see Table 1a).
- Single-population: if all samples belong to a single population without subpopulations, no change to the .fam file is required; provide a three-letter POP1 code (see Table 2a).
- Multi-population (“subpop”): if samples are labeled by subpopulation, the first column lists subpopulation codes and the user supply a three-letter POP1 tag separately (see Table 3a).

Coordinate file for mapping (required for pop1pop2 or subpop): tab-separated text (.txt) with five columns and no header: “POP1”,”POP2”,”Pop_name”,”LAT”,”LON”. Coordinates are taken as those of POP2; POP1 coordinates are computed internally as the mean of their corresponding POP2 locations (see Tables 1–3b).

By default, BPGA limits uploads to 600 MB (configurable), and subsets to autosomes when requested.

Genome builds & harmonization

BPGA accepts datasets aligned to hg18/GRCh36, hg19/GRCh37, or hg38/GRCh38. Inputs in hg18 or hg19 are lifted over to hg38 during preprocessing before any merge with the reference panel. During merging, variant alleles are harmonized to the reference panel's coding (including strand flips when resolvable); markers that cannot be aligned unambiguously (e.g., some A/T or C/G sites) are excluded.

Core analyses

1. PCA — LD-pruned variants (--indep-pairwise 200 25 0.3) are used to compute principal components via PLINK. Scatterplots can be colored by super- or subpopulation and exported as PNG.
2. *Fst* — Genome-wide *Fst* is computed with GCTA using user-selected comparisons. Results are shown as interactive Manhattan-style plots (plotly), with contextual links to dbSNP and the UCSC Genome Browser for clicked markers.
3. ADMIXTURE — Maximum-likelihood ancestry proportions are estimated for a user-selected single K or a range of K values, reporting cross-validation (CV) errors. BPGA renders stacked barplots and faceted boxplots and can display per-population mean ancestry proportions as geo-referenced pie charts (leaflet mini-charts).

Reference panels

BPGA optionally integrates user data with global reference datasets derived from the 1000 Genomes Project and the Human Genome Diversity Project [6,7]. A curated worldwide panel comprising 30,000 variants—with a genotyping rate of 0.997—was constructed from 3,258 individuals who passed quality control and filtering criteria. This panel serves as the reference. Subsetting and harmonization procedures ensure that only intersecting biallelic autosomal variants are retained. Users are advised to verify the redistribution terms applicable to any bundled subsets.

Use cases

Genotype data from HapMap Phase 3 were downloaded from the public repository (ftp: https://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/hapmap3/plink_format/draft_1/; file hapmap3_r1_b36_fwd.qc.poly.tar.bz2). The archive was decompressed, and PED/MAP files were converted to binary BED/BIM/FAM and filtered to MAF ≥ 0.05 with PLINK. A random set of 10,000 autosomal SNPs were extracted and used to filter full data set to obtain a final hapmap3_r1_b36 10K sample to test BPGA functionality. Fam file was formatted according BPGA data model as "pop1pop2" were POP1 were superpopulations (AFR=African; AMR=Admixed American; EAS=East Asian; EUR=European; SAS=South Asian) and POP2 subpopulations (ASW [African ancestry in Southwest USA], CEU [Utah residents with Northern and Western European ancestry from the CEPH collection], CHB [Han Chinese in Beijing, China], CHD [Chinese in Metropolitan Denver, Colorado], GIH [Gujarati Indians in Houston, Texas], JPT [Japanese in Tokyo, Japan], LWK [Luhya in Webuye, Kenya], MEX [Mexican ancestry in Los Angeles, California], MKK [Maasai in Kinyawa, Kenya], TSI [Toscans in Italy], YRI [Yoruba in Ibadan, Nigeria]). Geographic coordinates were assigned according to each population's geographic origin. Hapmap3_r1_b36 10K sample file was loaded to BPGA and processed according BPGA workflow (Table 4). Variant positions were lifted over to GRCh38 (hg38) to ensure consistent SNP coordinates for the interactive *Fst* plot.

Results

Figure 1 shows results obtained by BPGA analysis of the HapMap subset (MAF = 0.05; 9,326 variants, 898 individuals). The PC1–PC2 scatter cleanly recovers canonical continental structure—AFR separates maximally, EAS and EUR form compact clusters, SAS lies intermediate, and AMR shows the expected dispersion from mixed ancestry. A genome-wide CEU *vs*. YRI *Fst* scan reveals a low–to-moderate background with discrete high-*Fst* peaks at localized *loci*, consistent with continental divergence (top five *Fst* SNPs are listed on text window). ADMIXTURE at K = 4 (MAF = 0.05; CV = 0.520; 9,326 variants) resolves four continental components, quantifying two-way mixing in SAS and multi-way mixing in AMR; group-level boxplots and geographic pie maps recapitulate these patterns. Taken together, these concordant signals across PCA, *Fst*, and ADMIXTURE demonstrate that BPGA's default parameterization (MAF filtering and pruning as configured in the interface) reproduces well-established global structure while providing *locus*-specific (*Fst*) and individual-level (ADMIXTURE) resolution. The full BPGA workflow is summarized in Table 4.

In a typical laptop setting (4–8 cores), the example datasets complete within minutes; larger cohorts scale with available CPU/memory and disk I/O because PCA/ADMIXTURE use compiled binaries.

Discussion

Figure 1 provides an external benchmark for BPGA: with ~9k common SNPs (MAF ≥ 0.05), PCA cleanly recovers canonical continental structure; a CEU–YRI genome-wide *Fst* scan shows a low background punctuated by discrete high-*Fst* peaks; and ADMIXTURE at K=4 quantifies the expected two-way mixing in SAS and multi-way mixing in AMR. The concordance across genome-wide (PCA), locus-specific (*Fst*), and individual-level (ADMIXTURE) summaries indicates that BPGA's default parameterization (MAF filtering, LD pruning, CV-guided K) is sufficient for robust first-pass inference and QC. While HapMap ascertainment and the focus on common variants may understate fine-scale or rare-variant structure, the patterns align with established literature and support generalization to user datasets.

BPGA's design emphasizes transparency (plain-text logs, explicit parameters) and compatibility (delegating computation to PLINK, GCTA, ADMIXTURE rather than re-implementing algorithms). This makes it well suited to two complementary use modes. (i) Teaching/training: instructors can stage inquiry-based labs where learners vary MAF/LD pruning and observe effects on PCA; compare Fst profiles between subpopulations; and explore ADMIXTURE solutions across K. Georeferenced pie-charts connect statistical clusters to sampling geography while prompting discussion of ascertainment, labeling, and ethical interpretation. (ii) Prospective research: working groups can accelerate intake QC and cohort characterization (PCA clustering, outlier detection, preliminary ancestry assignment), run exploratory differentiation scans (Fst Manhattan plots with dbSNP/UCSC links for rapid locus triage), and rapidly prototype ADMIXTURE analyses to gauge plausible K and component stability before larger runs.

These analyses are intentionally exploratory. BPGA is not a substitute for full, scripted workflows; rather, it front-loads insight and produces parameterized outputs that can be reproduced and hardened in command-line pipelines.

Limitations. Run time scales with dataset size and hardware; cluster scheduling is not built-in. Model selection for ADMIXTURE (choice of K) still requires judgment balancing CV error and interpretability. Map visualizations rely on reasonable population coordinates. Fst values are sensitive to SNP density and filtering choices (beyond the LD pruning used for PCA). As always, ancestry inference should be framed cautiously to avoid over-interpretation or essentialist claims, and analyses must comply with data-use agreements and privacy regulations.

Ethics and data protection

BPGA processes genotypes that may be sensitive. The application does not require personally identifiable information beyond standard PLINK identifiers. Users must ensure that datasets are de-identified, ethically sourced, and used in accordance with local regulations and approved protocols.

Data and code availability

Source code is openly available at https://github.com/jfibla/BPGA-a-Shiny-app-to-perform-Basic-Population-Genetic-Analysis. Reference panel is derived from public datasets (1000 Genomes and HGDP). Example HapMap data set and coordinates file can be downloaded from https://doi.org/10.6084/m9.figshare.30011614.v1.

A hosted demo is also available at http://15.188.54.171:3838/bpga_app/. This server is meant for functionality testing and demos. Performance and availability may vary depending on server load. The demo uses the app's default upload cap (configured in the code) and includes example datasets to explore the workflow quickly.

Author contributions

Joan Fibla: Conceptualization, software, validation, visualization, documentation, writing—original draft; writing—review & editing.

Acknowledgments

I thank the developers and communities behind R, shiny, ggplot2, plotly, leaflet, PLINK, GCTA, ADMIXTURE, and the 1000 Genomes and HGDP consortia for foundational tools and resources.

Development and writing also benefited from assistance by ChatGPT (GPT-5) for code refactoring and editorial support; the author assumes full responsibility for the final content.

Funding

No dedicated funding was received for this work.

Competing interests

The author declares no competing interests.

References

[1] Purcell S, Neale B, Todd-Brown K, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81(3):559–575.

[2] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4:7.

[3] Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. American Journal of Human Genetics 88(1):76–82.

[4] Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Research 19(9):1655–1664.

[5] Chang W, Cheng J, Allaire JJ, Sievert C, Schloerke B, Xie Y, Allen J, et al. (2024) shiny: Web Application Framework for R. R package version 1.x.

[6] The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68–74.

[7] Bergström A, McCarthy SA, Hui R, et al. (2020) Insights into human genetic variation and population history from 929 diverse genomes. Science 367(6484): easay5012.

[8] International HapMap 3 Consortium; Altshuler DM, Gibbs RA, Peltonen L, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 2;467(7311):52-8.

Table 1a.- First column preformatted as POP1_POP2

Family ID	Individual ID	Father	Mother	Sex	Phenotype
AFR_ASW	NA19916	0	0	1	-9
AFR_ASW	NA19703	0	0	1	-9
EUR_CEU	NA12341	0	0	2	-9
EUR_CEU	NA06984	0	0	1	-9
EAS_CHB	NA18532	0	0	2	-9
EAS_CHB	NA18561	0	0	1	-9

Table 1b.-Assigned coordinate file (.txt)

POP1	POP2	Pop_name	LAT	LON
AFR	ASW	African	-3.82	12.93
EAS	CHD	Chinese	40.00	115.00
EUR	CEU	European	51.30	11.66

Note: Coordinates are from POP2 populations; POP1 coordinates are estimated as mean values of their subpopulations (POP2).

Table 2a.- First column describes one single population

Family ID	Individual ID	Father	Mother	Sex	Phenotype
AFR	NA19818	0	0	1	-9
AFR	NA20346	0	0	1	-9
AFR	NA19921	0	0	2	-9
AFR	NA20281	0	0	1	-9
AFR	NA20301	0	0	2	-9
AFR	NA20294	0	0	2	-9
AFR	NA20357	0	0	2	-9

Table 2b.- User-defined code- coordinates

User assigned population code
User assigned LON, LAT coordinates

Table 3a.- First column describes several populations

Family ID	Individual ID	Father	Mother	Sex	Phenotype
ASW	NA19916	0	0	1	-9
ASW	NA19703	0	0	1	-9
CEU	NA12341	0	0	2	-9
CEU	NA06984	0	0	1	-9
CHB	NA18532	0	0	2	-9
CHB	NA18561	0	0	1	-9

Table 3b.- Assigned coordinate file (.txt)

POP1	POP2	Pop_name	LAT	LON
ASW	ASW	Africa_sample	9.3	19.3
CEU	CEU	Europe_sample	50	15
CHB	CHB	Asia_sample	38	83

Note: Coordinates are from POP2 populations; POP1 coordinates are estimated as mean values of all (POP2) populations.

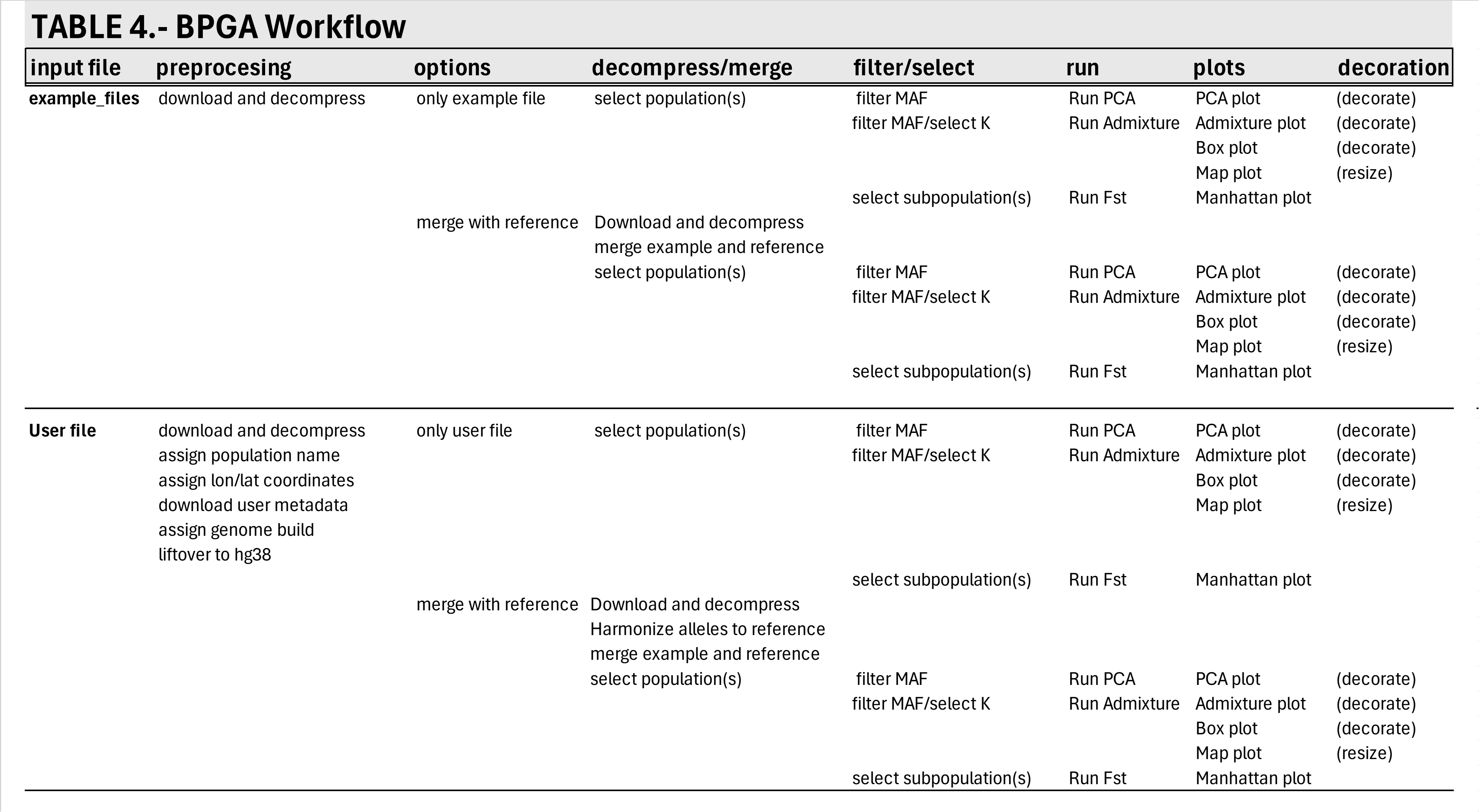
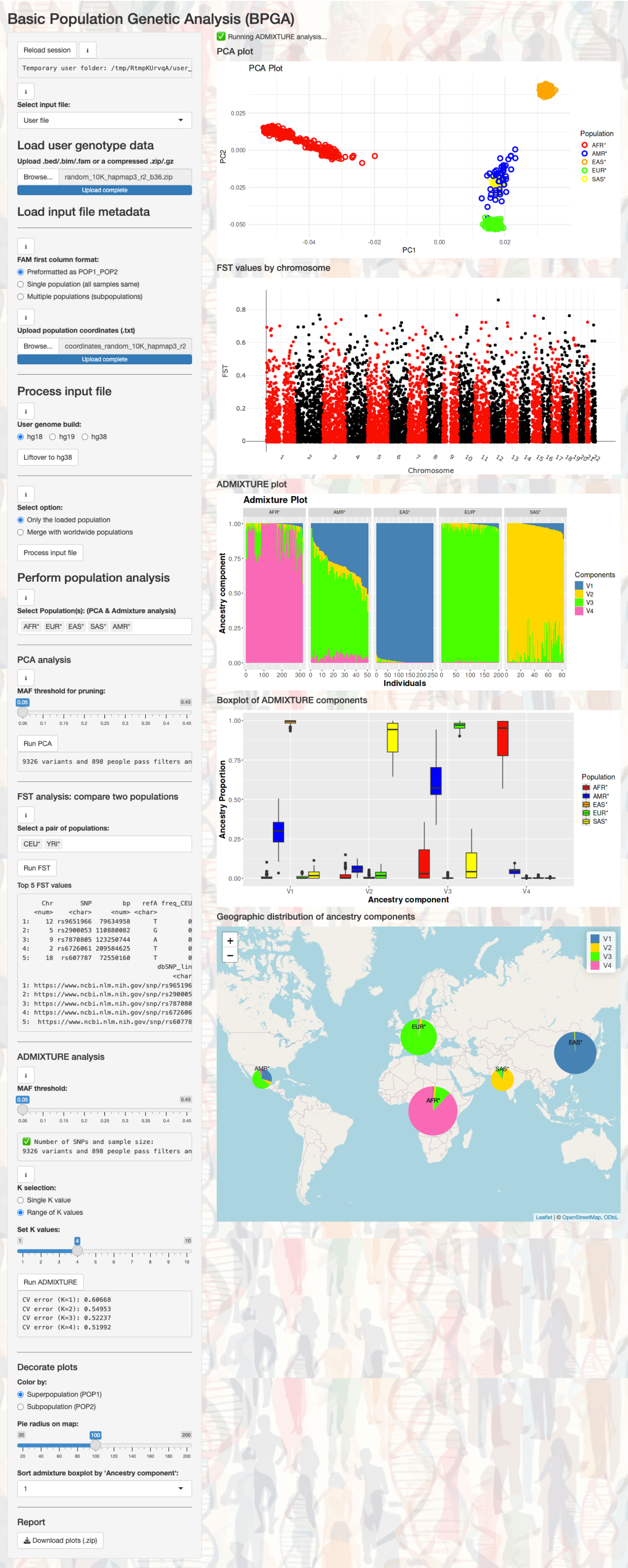


TABLE 4.- BPGA Workflow

input file	preprocesing	options	decompress/merge	filter/select	run	plots	decoration
example_files	download and decompress	only example file	select population(s)	filter MAF	Run PCA	PCA plot	(decorate)
				filter MAF/select K	Run Admixture	Admixture plot	(decorate)
						Box plot	(decorate)
						Map plot	(resize)
				select subpopulation(s)	Run Fst	Manhattan plot	
		merge with reference	Download and decompress				
			merge example and reference				
			select population(s)	filter MAF	Run PCA	PCA plot	(decorate)
				filter MAF/select K	Run Admixture	Admixture plot	(decorate)
						Box plot	(decorate)
						Map plot	(resize)
				select subpopulation(s)	Run Fst	Manhattan plot	
User file	download and decompress	only user file	select population(s)	filter MAF	Run PCA	PCA plot	(decorate)
	assign population name			filter MAF/select K	Run Admixture	Admixture plot	(decorate)
	assign lon/lat coordinates					Box plot	(decorate)
	download user metadata					Map plot	(resize)
	assign genome build						
	liftover to hg38						
				select subpopulation(s)	Run Fst	Manhattan plot	
		merge with reference	Download and decompress				
			Harmonize alleles to reference				
			merge example and reference				
			select population(s)	filter MAF	Run PCA	PCA plot	(decorate)
				filter MAF/select K	Run Admixture	Admixture plot	(decorate)
						Box plot	(decorate)
						Map plot	(resize)
				select subpopulation(s)	Run Fst	Manhattan plot	



Basic Population Genetic Analysis (BPGA)

Reload session

```
Temporary user folder: /tmp/RtmpKUrvqA/user_
```

Select input file: User file

Load user genotype data

Upload .bed/.bim/.fam or a compressed .zip/.gz

Browse... random_10K_hapmap3_r2_b36.zip — Upload complete

Load input file metadata

FAM first column format:

● Preformatted as POP1_POP2
○ Single population (all samples same)
○ Multiple populations (subpopulations)

Upload population coordinates (.txt)

Browse... coordinates_random_10K_hapmap3_r2 — Upload complete

Process input file

User genome build: ● hg18 ○ hg19 ○ hg38

Liftover to hg38

Select option:

● Only the loaded population
○ Merge with worldwide populations

Process input file

Perform population analysis

Select Population(s): (PCA & Admixture analysis) AFR* EUR* EAS* SAS* AMR*

PCA analysis

MAF threshold for pruning: 0.05

Run PCA

```
9326 variants and 898 people pass filters an
```

FST analysis: compare two populations

Select a pair of populations: CEU* YRI*

Run FST

Top 5 FST values

```
     Chr       SNP        bp   refA freq_CEU
   <num>    <char>     <num> <char>
1:    12 rs9651966  79634958      T        0
2:     5 rs2900053 110880082      G        0
3:     9 rs7870805 123250744      A        0
4:     2 rs6726061 209584625      T        0
5:    18  rs607787  72550160      T        0
                                      dbSNP_lin
                                          <char
1: https://www.ncbi.nlm.nih.gov/snp/rs965196
2: https://www.ncbi.nlm.nih.gov/snp/rs290005
3: https://www.ncbi.nlm.nih.gov/snp/rs787080
4: https://www.ncbi.nlm.nih.gov/snp/rs672606
5:  https://www.ncbi.nlm.nih.gov/snp/rs60778
```

ADMIXTURE analysis

MAF threshold: 0.05

```
✅ Number of SNPs and sample size:
9326 variants and 898 people pass filters an
```

K selection:

○ Single K value
● Range of K values

Set K values: 4

Run ADMIXTURE

```
CV error (K=1): 0.60668
CV error (K=2): 0.54953
CV error (K=3): 0.52237
CV error (K=4): 0.51992
```

Decorate plots

Color by:

● Superpopulation (POP1)
○ Subpopulation (POP2)

Pie radius on map: 100

Sort admixture boxplot by 'Ancestry component': 1

Report

Download plots (.zip)

✅ Running ADMIXTURE analysis...

***Figure 1**.* BPGA analysis of a random autosomal 10K-SNP HapMap sample. Image from top to bottom: (i) PCA scatterplot (PC1 vs PC2; MAF=0.05; 9,326 variants/898 subjects) shows AFR maximally separated on PC1, tight EAS and EUR clusters, SAS intermediate, and a dispersed AMR cloud; (ii) genome-wide FST (CEU vs YRI) reveals typical low–moderate differentiation punctuated by sharp high-FST peaks with top SNPs at local maxima; (iii) ADMIXTURE barplot (K=2-4; (CV error (K=1): 0.607, (K=2): 0.550 (K=3): 0.522, (K=4): 0.520) resolves four continental components with two-way mixing in SAS and multi-way mixing in AMR; (iv) boxplots summarize group-level ancestry (narrow AFR/EAS/EUR vs broader SAS/AMR); and (v) geographic pies map mean ancestries consistent with these patterns (AFR=African; AMR=Admixed American; EAS=East Asian; EUR=European; SAS=South Asian).